Estimating Large Delay Probabilities in Two Correlated Queues

EWAN JACOV CAHEN, CWI, Amsterdam
MICHEL MANDJES, University of Amsterdam
BERT ZWART, CWI, Amsterdam and Eindhoven University of Technology

This article focuses on evaluating the probability that both components of a two-dimensional stochastic process will ever, but not necessarily at the same time, exceed some large level u. An important application is in determining the probability of large delays occurring in two correlated queues. Since exact analysis of this probability seems prohibitive, we focus on deriving asymptotics and on developing efficient simulations techniques. Large deviations theory is used to characterise logarithmic asymptotics. The second part of this article focuses on efficient simulation techniques. Using "nearest-neighbour random walk" as an example, we first show that a "naive" implementation of importance sampling, based on the decay rate, is not asymptotically efficient. A different approach, which we call *partitioned* importance sampling, is developed and shown to be asymptotically efficient. The results are illustrated through various simulation experiments.

CCS Concepts: • **Mathematics of computing → Stochastic processes**; • **Networks → Network simulations**; • **Computing methodologies → Rare-event simulation**; • **General and reference →** *Estimation*; *Performance*;

Additional Key Words and Phrases: Importance sampling, large deviations, logarithmic asymptotics

ACM Reference format:
Ewan Jacov Cahen, Michel Mandjes, and Bert Zwart. 2018. Estimating Large Delay Probabilities in Two Correlated Queues. *ACM Trans. Model. Comput. Simul.* 28, 1, Article 2 (January 2018), 19 pages.
https://doi.org/10.1145/3158667

1 INTRODUCTION

Model. Consider a two-dimensional random walk $((A_s, B_s))_{s\in\mathbb{N}}$ with i.i.d. increments and with the partial sum processes denoted by

$$A_s := \sum_{i=1}^{s} X_i, \quad B_s := \sum_{i=1}^{s} Y_i.$$

The focus is on the probability $\pi(u) := \mathbb{P}\left(\exists s, t \in \mathbb{N} : A_s \geq u, \exists t \in \mathbb{N} : B_t \geq u\right)$, i.e., the probability of the event that both components will ever exceed some large level u but not necessarily at the same time. We allow that X_i and Y_i are dependent; note that if they were independent, the probability of interest would simply be the product of the marginal probabilities. An exact analysis of $\pi(u)$, however, seems possible only in special cases. In all of them, the model is such that

Authors' addresses: E. J. Cahen and B. Zwart, P.O. Box 94079, 1090 GB Amsterdam, NETHERLANDS; emails: {ewan.cahen, bert.zwart}@cwi.nl; M. Mandjes, POSTBUS 94248, 1090 GE Amsterdam, The Netherlands; email: m.r.h.mandjes@uva.nl.

https://doi.org/10.1145/3158667

the components are ordered: $A_s \leq B_s$ for all s, which implies that the epochs that the two components achieve their respective maximum values are almost surely ordered. These special cases cover tandem systems of $M/D/1$ queues [18] and a tandem Brownian queue [17]. In both models, let D_s denote the amount of work that has arrived up to time s and let c_1 and c_2 denote the constant service rate of the upstream and downstream queue, respectively (assume that $c_1 > c_2$). The stationary workloads of the upstream and the sum of the upstream and downstream queues are then respectively distributed as the supremum of $A_s = D_s - c_1 s$ and the supremum of $B_s = D_s - c_2 s$. In the above-mentioned articles, the analysis relies on the availability of the distribution of the maximum over a finite interval, given that we know the position at the end of the interval (in the Brownian case, this is a Brownian bridge; in the $M/D/1$ case, it can be dealt with using ballot theorems); the fact that such results are not generally available complicates the extension to more general models.

In the transform domain, results for the joint distribution of $\sup_{s\in\mathbb{N}} A_s$ and $\sup_{s\in\mathbb{N}} A_s$ have been established under more general conditions, but still an "ordering property" of the type mentioned above needs to be imposed; see, for instance, References [7, 14, 15]. Having expressions for such multivariate transforms, one still need to perform numerical inversion to obtain numerical output, which tends to be challenging in the tail of the multivariate distribution. Therefore, we resort in this article to large deviations and to rare-event simulation. We consider the rare-event regime in which both $\mathbb{E}(X_i) < 0$ and $\mathbb{E}(Y_i) < 0$.

This model has several applications. First, it can model two correlated queues. A queue is essentially a stochastic process reflected at zero. Consider now two queues fed by the (possibly correlated) input processes A_s and B_s. Then their stationary versions obey the distributional equalities, see, e.g., Reference [11, Section 1.1],

$$Q_1 \stackrel{\mathrm{d}}{=} \sup_s A_{-s}, \quad Q_2 \stackrel{\mathrm{d}}{=} \sup_s B_{-s},$$

which follow as a direct application from Lindley's recursion. The steady-state probability of both queues having more than an amount u of work is therefore equal to

$$\mathbb{P}(Q_1 > u, Q_2 > u) = \mathbb{P}(\exists s : A_{-s} > u, \exists t : B_{-t} > u),$$

which is, after reversing time, precisely the probability of our interest. Another application is in risk management, where we can use the model to study rare events in the context of two correlated portfolios; see, e.g., Reference [1] and references therein.

Literature. There is a substantial literature on efficient estimation of rare-event probabilities for queueing systems, see, e.g., the surveys in References [4], [12], and [13]. We here provide an account of this literature (without aiming at being exhaustive), focusing on multivariate rare events. In addition, we briefly comment on how these results relate to ours.

Over the past few decades, different techniques have been developed, the most prominent being importance sampling (based on a change of measure) and splitting; our present study falls in the former category. Building on the ideas of, e.g., Reference [20], Reference [19] focuses on estimating overflow-related quantities in a stable $GI/GI/m$ queue using importance sampling. Later attention shifted to more sophisticated queueing systems. In Reference [6] it is assessed to what extent state-independent change of measures can lead to asymptotically efficient performance in two-node tandem Jackson networks. The event of interest in that article is of an overflow probability in a two-node tandem Jackson network, whereas we focus on the two components of a two-dimensional random walk *both* ever reaching a high level. In Reference [10], where the focus is on two-node tandem Jackson networks, too, the authors consider, contrary to our article, state-*dependent* changes of measure. A generalisation to arbitrary Jackson networks is treated in

Reference [9]. In both articles, the so-called subsolution method is used, which is also briefly discussed in this article. Improved results are given in Reference [2], where the author focuses on optimal simulation algorithms for overflow probabilities during a busy period. Instead of using exponential twisting forward in time, the author proposes a method that goes backwards in time.

This article is a logical continuation of our previous work, see Reference [5]. In that article, we study a similar model, but there the event of interest corresponds to both components exceeding a large level *at the same time* (whereas in the present article these epochs can be different).

Decay Rate. The first result of this article is Theorem 3.1, which provides an expression for the *decay rate*,

$$\lim_{u\to\infty} \frac{1}{u} \ln \pi(u).$$

The proof of this result uses two important theorems in large deviations theory, namely Cramér's theorem and Mogulskii's theorem. In the proof, the lower bound is attained by conditioning where the "slower" component of the process is when the "fast" component hits level u for the first time. For the upper bound, we first show that the decay rate can be bounded by the decay rate of probability of the event of interest occurring on a bounded time interval. We then use this bounded interval to apply Mogulskii's theorem. Then we apply a "linear geodesics" type of argument to show that the obtained rate function over general sample paths is the same as over some set of piecewise linear sample paths.

Importance Sampling and Challenges. The second result of this article is the construction of an efficient simulation method to estimate $\pi(u)$. Since Monte Carlo simulation is slow due to the rarity of the event of interest when u gets large, we resort to importance sampling (IS). Importance sampling is a method to simulate stochastic systems using a different underlying probability measure, such that the (rare) event of interest is not rare any more; the simulation output is weighted by appropriate likelihood ratios to recover unbiasedness. Each probability measure leads to a particular variance performance, and it is therefore crucial to identify the one that is, according to some specific definition, optimal. There exist various performance metrics; the metric we use is called *asymptotic optimality*. We refer to Section 4 for the definition. For our IS procedure, we first propose a "naive" change of measure based on the decay rate given in Theorem 3.1. We will show, however, that this approach does not necessarily perform well. More specifically, we show that using this new measure for a "nearest-neighbour random walk" results in a procedure that is not asymptotically optimal. The underlying problem with this procedure is that at the moment when the "fast" component of the process hits level u, we do not have any control over the position of the "slow" component.

Partitioned IS. To solve this problem, we introduce *partitioned importance sampling*. This approach is based on conditioning where the "slow" component of the process has to be when the "fast" component hits level u for the first time. More specifically, we partition the event of interest into disjoint events and perform simulations to estimate the probabilities corresponding to those events. For more details, see Section 5. We show that this approach is indeed asymptotically optimal. It is pointed out how the method's inherent bias, arising from the need to truncate the infinite sum obtained through this method, can be made arbitrarily small.

Numerical Results. The results above are illustrated through various numerical experiments. To carry out the simulations, we chose a specific instance of the model, namely a model in which the increments (X_i, Y_i) have a bivariate normal distribution. We investigate how the performance of the three simulation methods described above (i.e., Monte Carlo, naive IS, and partitioned IS)

depend on various factors, e.g., the level u, the covariance of the two components, and the number of partitions used in partitioned importance sampling.

Organisation of the Paper. The rest of this article is organised as follows. Section 2 gives a detailed description of the model and a brief overview of large-deviations theory. In Section 3, we state the first main result of this article, namely an expression for the decay rate of the probability of the event of interest. In Section 4, we give a first naive importance sampling-based simulation scheme, and we show that this method is not asymptotically optimal. This is remedied in Section 5, where we introduce partitioned importance sampling. Moreover, we show that this new approach is indeed asymptotically optimal. These findings are illustrated in Section 6, where we give numerical results of various simulation experiments. The proof of the result in Section 3 is given in Section 7.

2 MODEL DESCRIPTION AND PRELIMINARIES

2.1 The Model

Consider the bivariate random walk $((A_s, B_s))_{s\in\mathbb{N}}$, where the partial sum process is denoted by, for $s \in \mathbb{N}$,

$$A_s := \sum_{i=1}^{s} X_i, \quad B_s := \sum_{i=1}^{s} Y_i,$$

with (X_i, Y_i) i.i.d. bivariate random vectors (whose components are not necessarily independent). We also introduce the events

$$\mathscr{A}_s(u) \equiv \mathscr{A}_s := \{A_s \geq u\}, \quad \mathscr{B}_s(u) \equiv \mathscr{B}_s := \{B_s \geq u\}.$$

The main object of study of this article is the probability $\pi(u)$ that both A and B exceed some (large) threshold u, but not necessarily at the same time:

$$\pi(u) := \mathbb{P}\left(\exists s \in \mathbb{N} : A_s \geq u, \exists t \in \mathbb{N} : B_t \geq u\right) = \mathbb{P}\left(\left(\bigcup_{s=1}^{\infty} \mathscr{A}_s\right) \cap \left(\bigcup_{t=1}^{\infty} \mathscr{B}_t\right)\right).$$

It is assumed throughout that both $\mathbb{E}(X_1)$ and $\mathbb{E}(Y_1)$ are negative, such that $\pi(u)$ is a rare-event probability, for u large.

Since an exact analysis of this probability seems not possible in general, we will look at the so-called *decay rate* of this probability:

$$\lim_{u\to\infty} \frac{1}{u} \ln \pi(u). \tag{1}$$

Our main result, which is an expression for the decay rate and which is stated in the next section, depends on both Cramér's and Mogulskii's theorem. Before we state Cramér's theorem below, a quick recap of some large deviation theory is first given. We follow the setup of Reference [11].

2.2 Preliminaries from Large Deviations

We define the *limiting cumulant generating function* of $((A_s, B_s))_{s\geq 0}$ as

$$\ln \Lambda(\theta, \eta) := \lim_{s\to\infty} \frac{1}{s} \ln \mathbb{E}\left(e^{\theta A_s + \eta B_s}\right) = \ln \mathbb{E}\left(e^{\theta X + \eta Y}\right). \tag{2}$$

A function $I : \mathbb{R}^2 \to \mathbb{R}^*$ (where $\mathbb{R}^* := \mathbb{R} \cup \{\infty\}$) is a *rate function* if it is non-negative and if it is lower semi-continuous, i.e., all level sets are closed (level sets of some function f are sets of the form $\{x : f(x) \leq \alpha\}, \alpha \in \mathbb{R}$). Furthermore, it is called a *good* rate function if in addition all level sets

are compact. We say that $((A_s, B_s))_{t\geq 0}$ satisfies a *large deviations principle* in $\mathbb{R}^2$ with rate function $I : \mathbb{R}^2 \to \mathbb{R}^*$ if for any measurable set $F \subseteq \mathbb{R}^2$

$$-\inf_{x\in F^\circ} I(x) \leq \liminf_{s\to\infty} \frac{1}{s} \ln \mathbb{P}\left((A_s, B_s) \in F\right) \leq \limsup_{s\to\infty} \frac{1}{s} \ln \mathbb{P}\left((A_s, B_s) \in F\right) \leq -\inf_{x\in F^c} I(x),$$

where F° and F^c denote the interior and closure of F, respectively. For any function $f : \mathbb{R}^d \to \mathbb{R}^*$, we denote its *convex conjugate* by $f^\star(x) := \sup_\theta \langle \theta, x\rangle - f(\theta)$.

3 LARGE DEVIATIONS RESULT

This section presents the first main result of the article, namely Theorem 3.1, which provides an expression for the decay rate (1). This result is based on both Cramér's theorem and Mogulskii's theorem. It requires the definition of joint (bivariate, that is) moment-generating functions and Legendre transforms. Recall that

$$\Lambda(\theta, \eta) = \mathbb{E}\left(e^{\theta X + \eta Y}\right);$$

we define its univariate counterparts through

$$\Lambda_1(\zeta) := \Lambda(\zeta, 0), \qquad \Lambda_2(\zeta) := \Lambda(0, \zeta),$$

which we assume to satisfy the condition of Mogulskii's theorem, i.e., we assume Equation (2) to be finite everywhere. In addition, the bivariate Legendre transform is given through

$$I(x, y) := \sup_{\theta,\eta} \left(\theta x + \eta y - \ln \Lambda(\theta, \eta)\right)$$

and its univariate counterparts through

$$I_1(x) := \sup_\zeta \left(\zeta x - \ln \Lambda_1(\zeta)\right), \qquad I_2(x) := \sup_\zeta \left(\zeta x - \ln \Lambda_2(\zeta)\right).$$

Define, for $i = 1, 2$,

$$\alpha_i := \inf_{z>0} \frac{I_i(z)}{z}. \tag{3}$$

The main result of this section gives an expression for the decay rate (1) in terms of a variational problem.

Theorem 3.1. *If (2) exists and is finite everywhere, then*

$$\begin{aligned}&\lim_{u\to\infty} \frac{1}{u} \ln \pi(u).\\ &= -\min\left\{\inf_{x>0, y\leq x}\left(\frac{I(x,y)}{x} + \left(1 - \frac{y}{x}\right)\alpha_2\right), \inf_{y>0, x<y}\left(\frac{I(x,y)}{y} + \left(1 - \frac{x}{y}\right)\alpha_1\right)\right\}.\end{aligned} \tag{4}$$

Proof. Due to the complexity and length, the proof is postponed to Section 7. Instead, we will give a short summary of the main ideas here.

We will prove this as a lower and an upper bound. For the lower bound, we first look at specific times s and t such that $A_s \geq u$ and $B_t \geq u$. We condition on where the slower process is, and call this position y, when the faster process hits level u. We then first use Cramér's theorem and subsequently optimise over the position y. Taking the supremum over s and t then gives us, after some rewriting, the correct decay rate. For the upper bound, we use the union bound to again condition where the slow process is when the fast process hits level u. We then argue that we only need to look at the probability of the event occurring in a bounded time interval. This allows us to use Mogulskii's theorem to arrive at the claimed expression. □

In the next section, we will focus on estimating $\pi(u)$ numerically for u large. In that section, we propose an implementation of importance sampling that is based on the ideas behind the decay

rate presented in Theorem 3.1. We will show, however, that this method, although natural, is not necessarily asymptotically efficient.

4 IMPORTANCE SAMPLING AND EFFICIENCY

In this and the next sections, we focus on estimating $\pi(u)$ numerically. We will show that a naive importance sampling procedure, based on the decay rate given by Theorem 3.1, is not asymptotically optimal in general; recall that a simulation procedure is called *asymptotically optimal* if, in self-evident notation,

$$\lim_{u\to\infty} \frac{\ln \mathbb{E}_{\mathbb{Q},u}\left(L^2 I\right)}{\ln \mathbb{E}_{\mathbb{Q},u}\left(LI\right)} = 2, \tag{5}$$

where $\mathbb{Q}$ is the new measure and L is the likelihood ratio, or Radon-Nikodym derivative, between $\mathbb{P}$ and $\mathbb{Q}$, i.e., $L = \frac{d\mathbb{P}}{d\mathbb{Q}}$. We refer to, e.g., Reference [4, Definition 1]) for background information on this optimality concept or Reference [16] for an in-depth account of various performance metrics. Note that by Jensen's inequality, the limit in Equation (5) is always smaller than or equal to 2, so it is left to prove that it is larger than or equal to 2. We say that we *exponentially twist* a random variable X, having density $f_{\mathbb{P}}(\cdot)$, with parameter θ if under $\mathbb{Q}$, the density of X equals (in self-evident notation) $f_{\mathbb{Q}}(x) = f_{\mathbb{P}}(x)e^{\theta x}/\mathbb{E}_{\mathbb{P}}(e^{\theta X})$.

As an instance of the model, we will consider a "nearest-neighbour random walk." More specifically, we let

$$(X_i, Y_i) = \begin{cases} (1,0) & \text{w.p. } p_1; \\ (1,1) & \text{w.p. } p_2; \\ (-1,-1) & \text{w.p. } p_3; \\ (-1,1) & \text{w.p. } p_4. \end{cases} \tag{6}$$

We will assume that the probabilities add up to unity.

The naive importance sampling procedure, which will yield the change of measure $\mathbb{Q}$, is as follows. Each simulation run consist of (up to) two sequential exponential twists; the first twist is used to bring one of the components up lo level u, whereas the second twist (if still necessary) is used to bring the other (slower) component up to level u. Denote by $\theta^\star$ and $\eta^\star$ the optimising parameters of $I(\cdot,\cdot)$ in Equation (4). We first exponentially twist the joint process with parameter $(\theta^\star, \eta^\star)$ until one of the components hits level u. If process A (B) hits level u first, then we exponentially twist process B (A) with parameter $\zeta^\star$, which is defined as the optimising parameter of $I_2(\cdot)$ ($I_1(\cdot)$), until this component hits level u. The following will be used in the proofs of Property 2 and Theorem 5.1: Let $S \equiv S(u)$ be the first passage time of level u for process A, i.e., $S := \inf\{s : A_s \geq u\}$. Furthermore, let T be the analogous counterpart for process B.

This procedure follows naturally from Theorem 3.1 as it tries to mimic the most likely path given in the theorem: First twist both processes until the fastest reaches level u and then twist the slower process.

Property 1. *For any x and y in the infimum of Theorem 3.1 (even the non-optimal values), if we perform the exponential twist as described above using the optimal $\theta^\star, \eta^\star$ corresponding to these x and y, then*

$$\mathbb{E}_{\mathbb{Q}}(X_i) = x, \qquad \mathbb{E}_{\mathbb{Q}}(Y_i) = y.$$

Proof. We only prove this for $\mathbb{E}_{\mathbb{Q}}(X_i)$, since $\mathbb{E}_{\mathbb{Q}}(Y_i)$ can be dealt with analogously. In the supremum of I, the first-order conditions are

$$x - \frac{\partial \ln \Lambda(\theta,\eta)}{\partial\theta} = x - \frac{\frac{\partial\Lambda(\theta,\eta)}{\partial\theta}}{\Lambda(\theta,\eta)} = 0, \qquad y - \frac{\partial \ln \Lambda(\theta,\eta)}{\partial\eta} = y - \frac{\frac{\partial\Lambda(\theta,\eta)}{\partial\eta}}{\Lambda(\theta,\eta)} = 0.$$

Note now that by twisting, we have

$$\mathbb{E}_{\mathbb{Q}}(X_i) = \frac{\int z e^{\theta^\star z} f(z)\,\mathrm{d}z}{\int e^{\theta^\star z} f(z)\,\mathrm{d}z} = \frac{\frac{\partial \Lambda(\theta,\eta)}{\partial \theta}}{\Lambda(\theta,\eta)} = x,$$

where the first equality comes from the definition of an exponential twist, the second equality comes from the definition of the moment generating function, and the final identity comes from the display above.

Under any exponential twist, the model will be as follows:

$$(X_i, Y_i) = \begin{cases} (1,0) & \text{w.p. } \tilde{p}_1; \\ (1,1) & \text{w.p. } \tilde{p}_2; \\ (-1,-1) & \text{w.p. } \tilde{p}_3; \\ (-1,1) & \text{w.p. } \tilde{p}_4. \end{cases} \tag{7}$$

Property 2. *The naive importance sampling procedure (using probability measure $\mathbb{Q}$) as described above is* not *asymptotically optimal.*

Proof. Without loss of generality assume that A hits level u first. In the case that B hits level u first, the proof is analogous. The proof indicates that the spread in the vertical direction, i.e., the position of B_S, at time S causes the procedure to be not asymptotically optimal. Realise that we can write

$$B_{S(u)} = \sum_{i=1}^{u} V_i,$$

where the V_i are i.i.d. and $V_i \overset{\mathrm{d}}{=} V$ corresponding to the vertical position at the moment the horizontal position (i.e., corresponding to A) attains the value 1 for the first time. Hence,

$$\mathbb{E}(z^{B_{S(u)}}) = (\mathbb{E}(z^V))^u = \phi(z)^u,$$

where the right-hand side can be seen as the uth power of some probability generating function. From this it follows that, for all $\theta > 0$,

$$\lim_{u\to\infty} \frac{1}{u} \ln \mathbb{E}(e^{\theta B_{S(u)}}) = \ln \mathbb{E}(e^{\theta V}) = \theta \mathbb{E}(V) + G(\theta), \tag{8}$$

with $G(\theta) := \ln \mathbb{E}(e^{\theta(V - \mathbb{E}(V))})$ such that $G(0) = G'(0) = 0$ and $G(\cdot)$ is strictly convex.

Note that the likelihood ratio can be written as

$$L(u) = \Lambda(\theta^\star, \eta^\star) \exp\left(-\theta^\star A_S - \eta^\star B_S\right) \cdot \Lambda(0, \zeta^\star) \exp\left(-\zeta^\star (B_T - B_S)\right).$$

It is an elementary exercise to verify that $\Lambda(\theta^\star, \eta^\star) = \Lambda(0, \zeta^\star) = 1$ (which can be checked, e.g., by working out the first-order conditions). Moreover, in this model we know that $A_S = B_T = u$. Hence, the likelihood ratio reads

$$L(u) = \exp\left(-\theta^\star u - \eta^\star B_S\right) \cdot \exp\left(-\zeta^\star (u - B_S)\right).$$

From this, we obtain

$$\lim_{u\to\infty} \frac{1}{u} \ln \mathbb{E}_{\mathbb{Q},u}(LI) = -(\theta^\star + \zeta^\star) + \lim_{u\to\infty} \frac{1}{u} \ln \mathbb{E}_{\mathbb{Q},u}\left(\exp(-(\eta^\star - \zeta^\star) B_S)\right),$$

and, in the same way, we get for the second moment that

$$\lim_{u\to\infty} \frac{1}{u} \ln \mathbb{E}_{\mathbb{Q},u}(L^2 I) = -2(\theta^\star + \zeta^\star) + \lim_{u\to\infty} \frac{1}{u} \ln \mathbb{E}_{\mathbb{Q},u}\left(\exp(-2(\eta^\star - \zeta^\star) B_S)\right).$$

From the above properties of $G(\cdot)$, it follows that $G(2\theta) > 2G(\theta)$ for all $\theta \neq 0$. From this, it follows that

$$\ln \mathbb{E}_{\mathbb{Q},u}\left(e^{-2(\eta^\star-\zeta^\star)(V-\mathbb{E}(V))}\right) > 2\ln \mathbb{E}_{\mathbb{Q},u}\left(e^{-(\eta^\star-\zeta^\star)(V-\mathbb{E}(V))}\right),$$

or, equivalently,

$$\begin{aligned} &-2(\theta^\star+\zeta^\star) + \lim_{u\to\infty}\frac{1}{u}\ln \mathbb{E}_{\mathbb{Q},u}\left(\exp(-2(\eta^\star-\zeta^\star)B_S)\right) \\ &> -2(\theta^\star+\zeta^\star) + 2\lim_{u\to\infty}\frac{1}{u}\ln \mathbb{E}_{\mathbb{Q},u}\left(\exp(-(\eta^\star-\zeta^\star)B_S)\right). \end{aligned}$$

This reduces to

$$\lim_{u\to\infty}\frac{1}{u}\ln \mathbb{E}_{\mathbb{Q},u}(L^2 I) > 2\lim_{u\to\infty}\frac{1}{u}\ln \mathbb{E}_{\mathbb{Q},u}(LI),$$

which indeed proves that the procedure is *not* asymptotically optimal. □

The proof above indicates that the naive importance sampling procedure cannot be guaranteed to be asymptotically efficient because of the spread in the position of the slower component when the faster component hits level u for the first time. In particular, there is no lower bound for this position. In the next section, we try to overcome this complication by introducing a method called "partitioned importance sampling." This is a method that the required control over the position of the slower process.

5 PARTITIONED IMPORTANCE SAMPLING

In the previous section, we have seen that ordinary importance sampling does not yield an asymptotically optimal procedure in general. In this section, we give a procedure that *is* asymptotically optimal, namely partitioned importance sampling. The method's inherent bias can be made arbitrarily small.

The procedure consists of intersecting the set of interest by a partitioning in terms of the value of B_S. More concretely, we consider a decomposition into probabilities of disjoint events. First, let

$$\begin{aligned} \pi_1(u) &:= \mathbb{P}\left(\exists s,t\in\mathbb{N}, s\le t : \forall r < s : A_s \ge u, B_r < u, B_t \ge u\right), \\ \pi_2(u) &:= \mathbb{P}\left(\exists s,t\in\mathbb{N}, s > t : \forall r \le t : B_t \ge u, A_r < u, A_s \ge u\right), \end{aligned}$$

i.e., $\pi_1(u)$ is the probability that A hits level u before B, and $\pi_2(u)$ is its analogous counterpart. Note that $\pi(u) = \pi_1(u) + \pi_2(u)$. It is sufficient to show how $\pi_1(u)$ can be estimated efficiently, since the method for estimating $\pi_2(u)$ can be set up analogously. The decomposition we consider is

$$\pi_1(u) = \sum_{k=-\infty}^{\infty} \pi_{1,k}(u),$$

where the probabilities $\pi_{1,k}(u)$ are defined by

$$\pi_{1,k}(u) := \mathbb{P}\left(\exists s,t\in\mathbb{N}, t\ge s : \forall r < s : A_s \ge u, B_r < u, B_t \ge u, B_s \in s_k\right),$$

with $s_k := [kf(u), (k+1)f(u))$, where $f(\cdot)$ is a positive function (on which we impose some conditions below). We let $m \equiv m(u), M \equiv M(u)$ be a suitably chosen truncation, possibly dependent on u, and hence we estimate

$$\pi_1^{(\mathrm{app})}(u) := \sum_{k=m}^{M} \pi_{1,k}(u);$$

clearly, by choosing m sufficiently small and M sufficiently large the error made is negligible. We furthermore, to guarantee asymptotic optimality, need to impose that $M(u) - m(u)$ grows subexponentially as a function of u, i.e., we require that $\lim_{u\to\infty}\frac{1}{u}\ln(M(u)-m(u)) = 0$. We also require

that $kf(u)/u \leq 1$ for each $k \in \{m(u), \ldots, M(u)\}$ and impose the related property that $f(u)$ grows sublinearly in u for reasons that will become clear soon. In our simulation procedure, we perform a separate simulation run for each $k \in \{m(u), \ldots, M(u)\}$ as follows.

We start by solving, for each $k \in \{m, \ldots, M\}$, using definition (3),

$$J_k(u) = \inf_x \frac{I(x, x\,k\frac{f(u)}{u})}{x} + \left(1 - k\frac{f(u)}{u}\right)^+ \alpha_2; \tag{9}$$

bear in mind that if $f(u)$ would have been allowed to grow superlinearly, then the second term would be identical 0 eventually (for positive k). Denote the optimisers by $\theta_k^\star$ and $\eta_k^\star$ (from the definition of $I(\cdot,\cdot)$) and $\zeta \equiv \zeta_k^\star$ (from the definition of $I_2(\cdot)$; use Equation (3) and conclude that this twist does not depend on k). Observe that the first element in the minimum of Equation (4) can now be majorised as follows: Equation (9), and in particular

$$J^{(1)} := \inf_{x>0, y\leq x} \left(\frac{I(x,y)}{x} + \left(1 - \frac{y}{x}\right)\alpha_2\right) \leq \lim_{u\to\infty} \inf_k J_k(u);$$

the reason for the inequality is that in the left-hand side the minimum is taken over a larger set than in the right-hand side. In our refined algorithm, when estimating $\pi_{1,k}(u)$, we first twist the (X_s, Y_s) by $(\theta_k^\star, \eta_k^\star)$ until A exceeds u, and from that point onward twist the (Y_s) by $\zeta^\star$ until B exceeds u (if needed). The simulation output of a single run is $L_k I_k$, with $L_k \equiv L_k(u)$ again the likelihood ratio, and the indicator function $I_k \equiv I_k(u)$, which equals 1 iff the path is such that both A and B exceed u but now in addition that (i) A is required to exceed u before B does (or simultaneously), and (ii) when A exceeds u, B is in the interval s_k; it is this latter requirement that gives us control on the variance of the estimator, as will be shown below.

Theorem 5.1. *The simulation procedure described above is asymptotically optimal.*

Proof. Observe that the likelihood reads

$$L_k(u) = \left(\Lambda(\theta_k^\star, \eta_k^\star)\right)^S \exp\left(-\theta_k^\star A_S - \eta_k^\star B_S\right) \cdot \left(\Lambda(0, \zeta_k^\star)\right)^{T-S} \exp\left(-\zeta^\star(B_T - B_S)\right).$$

As in the proof of Property 2, we have $\Lambda(\theta_k^\star, \eta_k^\star) = \Lambda(0, \zeta^\star) = 1$, so that we can simplify $L_k(u)$ to

$$L_k(u) = \exp\left(-\theta_k^\star A_S - \eta_k^\star B_S\right) \cdot \exp\left(-\zeta^\star(B_T - B_S)\right).$$

Now note that, on the event that $I_k(u) = 1$, it holds that (i) $A_S \geq u$, (ii) $B_S \in s_k$, and (iii) $B_T - B_S \geq (1 - (k+1)\,f(u)/u) \cdot u \geq 0$. Therefore, uniformly in u, using the definition of $J_k(u)$ and $\Lambda(\theta_k^\star, \eta_k^\star) = \Lambda(0, \zeta^\star) = 1$,

$$\begin{aligned}
\frac{1}{u} \ln L_k(u) I_k(u) &\leq -\theta_k^\star - \min\left\{\eta_k^\star\, k\frac{f(u)}{u}, \eta_k^\star\,(k+1)\frac{f(u)}{u}\right\} - \zeta^\star\left(1 - k\frac{f(u)}{u}\right)^+ + \zeta^\star\frac{f(u)}{u} \\
&\leq -\theta_k^\star - \eta_k^\star\, k\frac{f(u)}{u} - \zeta^\star\left(1 - k\frac{f(u)}{u}\right)^+ + \zeta^\star\frac{f(u)}{u} \\
&= -J_k(u) + \zeta^\star\frac{f(u)}{u};
\end{aligned}$$

the final equality in the above display can be seen by realising that

$$\alpha_2 = \inf_{z>0}\frac{I_i(z)}{z} = \inf_{z>0}\frac{\sup_\zeta \zeta z - \ln\Lambda_2(\zeta)}{z} = \frac{\zeta^\star z^\star - \ln\Lambda_2(\zeta^\star)}{z^\star} = \frac{\zeta^\star z^\star - 0}{z^\star} = \zeta^\star$$

and similarly for the first term.

We estimate our probability by evaluating sample averages of independent random variables that are distributed as

$$Z_{m,M}(u) := \sum_{k=m}^{M} L_k(u) I_k(u),$$

with m, M a suitably chosen truncation. Note that $Z_{m,M}(u)$ is bounded from above by

$$Z_{m,M}(u) \leq (M - m + 1) \exp\left(- \min_{k \in \{m,\ldots,M\}} J_k(u)u + \zeta^\star f(u) \right).$$

We thus see that, using that $M - m$ grows subexponentially and using that $f(u)$ is sublinear,

$$\limsup_{u\to\infty} \frac{1}{u} \ln \mathbb{E}\left((Z_{m,M}(u))^2 \right) \leq -2J^{(1)} = 2 \lim_{u\to\infty} \frac{1}{u} \ln \pi_1(u),$$

so the procedure is asymptotically optimal.

6 NUMERICAL RESULTS

In this section, we show and discuss several numerical experiments. Throughout this section, we use a specific instance of the general model as described in Section 2, which, in particular, satisfies the conditions of Theorem 3.1.

6.1 The Model

In the examples considered, we assume that the (X_i, Y_i) are i.i.d. vectors with a bivariate normal distribution with mean vector and covariance matrix given by

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} (\sigma_1)^2 & \rho \\ \rho & (\sigma_2)^2 \end{pmatrix},$$

respectively. One of the reasons that we chose this model is the following nice property.

Property 3. *If we exponentially twist (X_i, Y_i) with parameter (θ_1, θ_2), then the twisted process again has a bivariate normal distribution with mean vector*

$$\tilde{\mu} = \begin{pmatrix} \mu_1 + \theta_1(\sigma_1)^2 + \theta_2\rho \\ \mu_2 + \theta_1\rho + \theta_2(\sigma_2)^2 \end{pmatrix}$$

and covariance matrix $\tilde{\Sigma} = \Sigma$.

Proof. This follows from elementary calculations. □

Below we perform a number of different experiments, in which we test the influence of various parameters on the performance of the simulations. The simulations were carried out in R. Anytime in this section we refer to (non-partitioned) IS, *naive* importance sampling is meant. Unless otherwise stated, in all numerical experiments, we ran simulations until a 95% confidence interval with 10% precision was obtained. We tested both the number of runs and the running time (cpu time) required to obtain the confidence interval. In most cases, the cpu time shows the same quantitative behaviour as the number of runs. Therefore, the cpu times are only shown in Section 6.4, where this is not the case.

Remark. As already stated before, the truncation used in partitioned importance sampling inherently produces a biased estimator. It is, however, possible to obtain an estimator with *vanishing relative bias* as $u \to \infty$. This can be accomplished by choosing the lower bound $m(u)f(u)$ and upper bound $M(u)f(u)$ such that the expected value of the slower process, at the moment the faster

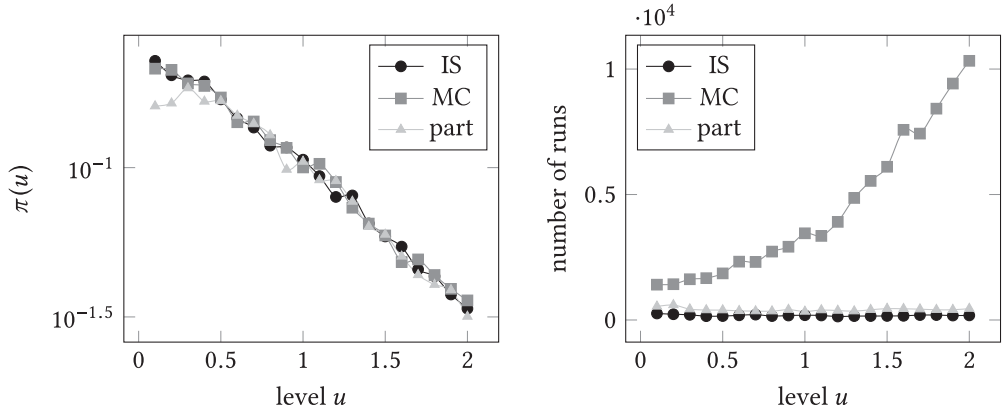

Fig. 1. These plots show the probability $\pi(u)$ and the number of runs needed to get the desired accuracy, respectively. The parameters that were used were $\mu = \begin{pmatrix} -1 \\ -0.5 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, $m(u) = -40u$, $M(u) = 20u - 1$, and $f(u) = 0.05$. The results show that importance sampling gives a significant efficiency improvement over Monte Carlo sampling.

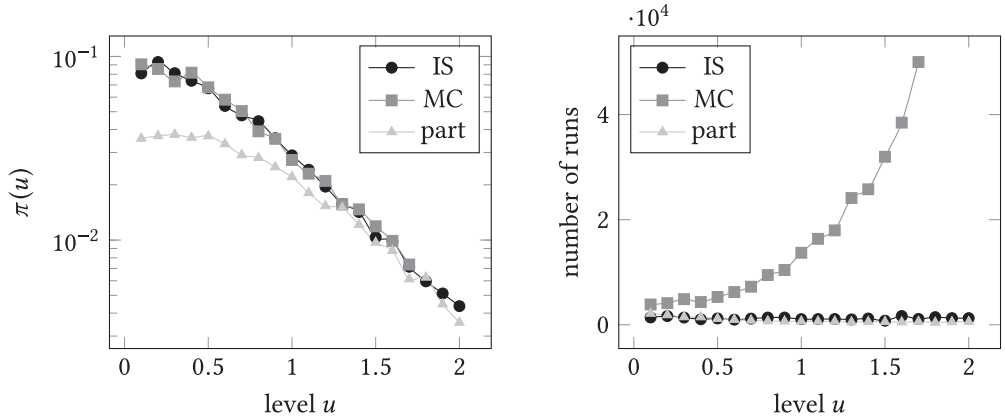

Fig. 2. These plots show the probability $\pi(u)$ and the number of runs needed to get the desired accuracy, respectively. We stopped simulations when 50,000 runs were needed, hence the missing values for Monte Carlo sampling. The parameters that were used were $\mu = \begin{pmatrix} -1 \\ -0.5 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$, $m(u) = -40u$, $M(u) = 20u - 1$, and $f(u) = 0.05$. The results show that importance sampling gives a signifivant efficiency improvement over Monte Carlo sampling.

process hits level u, is in between these bounds. This expected value can be numerically determined by combining Theorem 3.1 and Property 1. We indeed chose m, M, and f such that in the experiments in Figures 1, 2, and 3 we have vanishing relative bias.

Remark. There exist more methods of sampling than naive and partitioned importance sampling. One method makes use of a so-called *subsolution method*, see, e.g., Reference [9]. We implemented such a state-dependent importance sampling scheme in the following way. In Example 3 of Reference [3, pp. 47–48], it is explained how to set up a subsolution-based scheme for estimating the probability that *at least one of the components* reaches a rare set. The procedure is then as follows:

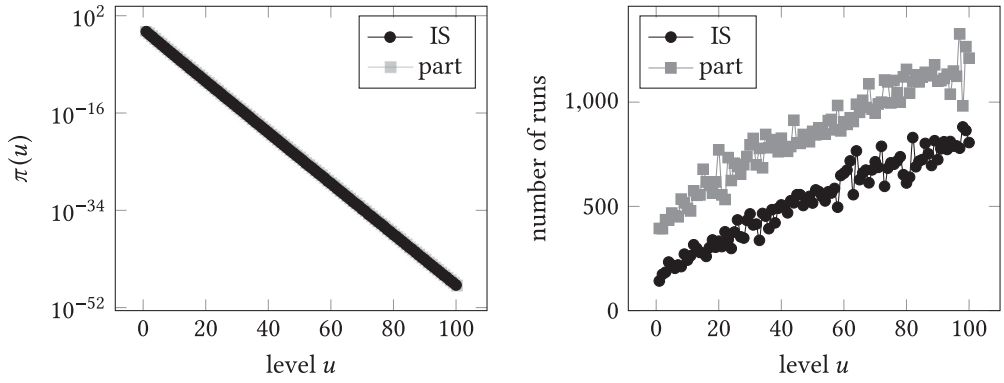

Fig. 3. These plots show the probability $\pi(u)$ and the number of runs needed to get the desired accuracy, respectively. The parameters that were used were $\mu = \binom{-1}{-0.5}$, $\Sigma = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, $m(u) = -10$, $M(u) = u - 1$, and $f(u) = 1$. The results show that for extremely small probabilities, the number of runs needed for IS and partitioned IS is small.

- In each simulation run, we used this scheme until one of the components exceeds level u;
- subsequently, we use a single exponential twist for the remainder of the run (with the relevant α_i, as given in Equation (3)), until the other component has exceeded level u as well.

We compared this method with the partitioned importance sampling method proposed in our article. We tested both procedures extensively, in terms of the number of runs and the cpu time. In the simulations we performed, we observed that the partitioned IS method performs better. We suspect that this may be due to the same problem as in the naïve IS scheme, namely the random fluctuations of the "second component" (for instance, the fluctuations of the vertical component at the epoch that the horizontal component first exceeds u)—the way we set up the subsolution-based scheme the fluctuations of the second component are apparently not sufficiently controlled (as in the first part of the run the focus is only on the event that one of the two components exceeds u).

Clearly, the partitioned importance sampling will become less attractive when considering problems of higher dimensions. Considering the counterpart of our problem but then in dimensions higher than 2, one could again come up with a partitioning such that $Z_k(u)$ can be written as sum (over all k) of $L_k(u)I_k(u)$, but the number of k to be included will increase (which will slow down the simulation). It is therefore anticipated that in higher dimensions subsolution-based schemes will become advantageous.

6.2 Variable Level

In this section, we look at how the level to reach u influences both the probability $\pi(u)$ and the number of simulations needed. We ran two different experiments of which the results can be found in Figures 1 and 2. The experiments differ in the sign of the covariance that was used; we refer to the respective caption for specific details. The results below clearly show that Monte Carlo sampling is much slower than both importance sampling and partitioned importance sampling. Furthermore, when comparing Figures 1 and 2, we see that a negative correlation between the two components negatively influence both the probability $\pi(u)$ and the number of samples needed to get the desired precision.

It should be noted that the event of interest in the experiments as described shown in Figures 1 and 2 above can hardly be called "rare." The reason that we kept the level to reach u relatively

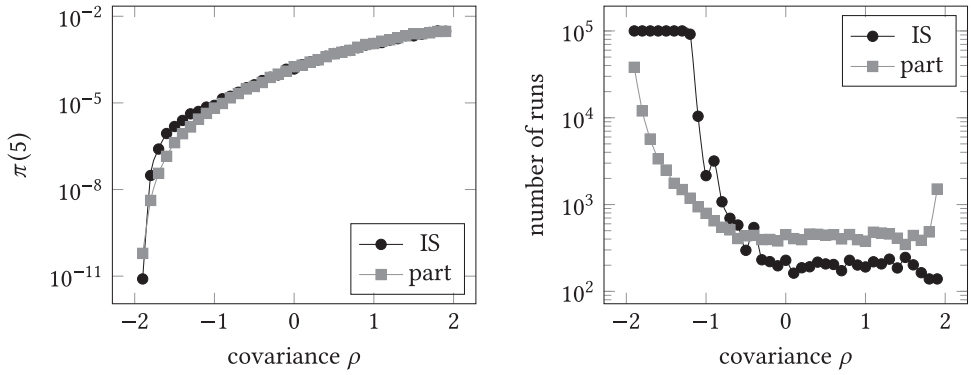

Fig. 4. These plots show the probability $\pi(5)$ and the number of runs (with a maximum of 100,000) needed to get the desired accuracy, respectively. The parameters that were used were $\mu = \begin{pmatrix} -1 \\ -0.5 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 2 & \rho \\ \rho & 2 \end{pmatrix}$. We fixed the number of intervals to 10, the lower bound of those intervals to −5, and the upper bound to 5. The results show that when the two components are strongly negative correlated, partitioned importance sampling behaves much better than ordinary importance sampling. When the components are positively correlated, the opposite holds.

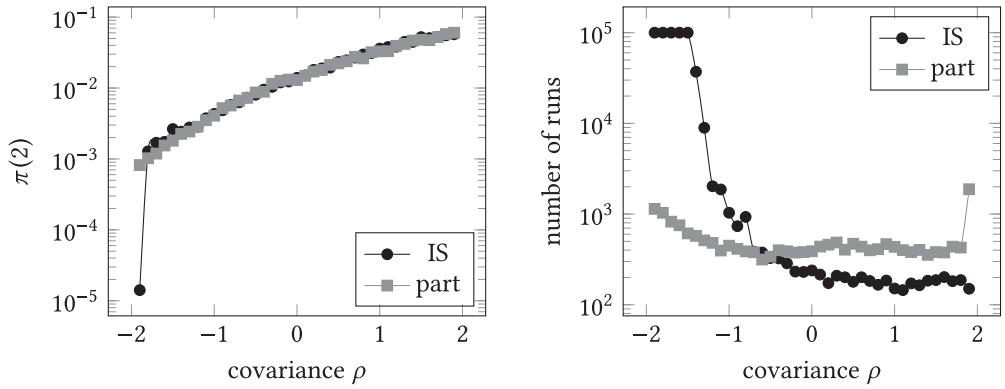

Fig. 5. The precise same experiment as reported in Figure 4, except that $u = 2$ now and the upper bound for the intervals also equals 2.

low is that Monte Carlo sampling quickly took over 10,000 runs, which takes a long time in R. Therefore, in Figure 3 we performed some simulations for bigger u, though only for importance sampling and partitioned importance sampling. The results clearly show that even for extremely small probabilities (around 10^{-52}), both importance sampling and partitioned importance sampling need a modest amount of runs.

6.3 Variable Covariance

Having convinced ourselves that Monte Carlo sampling is prohibitively slow, we will restrict the experiments now to only importance sampling and partitioned importance sampling. In the previous experiments, both IS and partitioned IS performed roughly the same: The number of runs required to get the desired confidence interval did not show any significant differences. We will now identify cases where partitioned IS behaves much better than IS. In the next simulations, we look at how the covariance influences both $\pi(u)$ and the number of trials needed. Figures 4 and 5

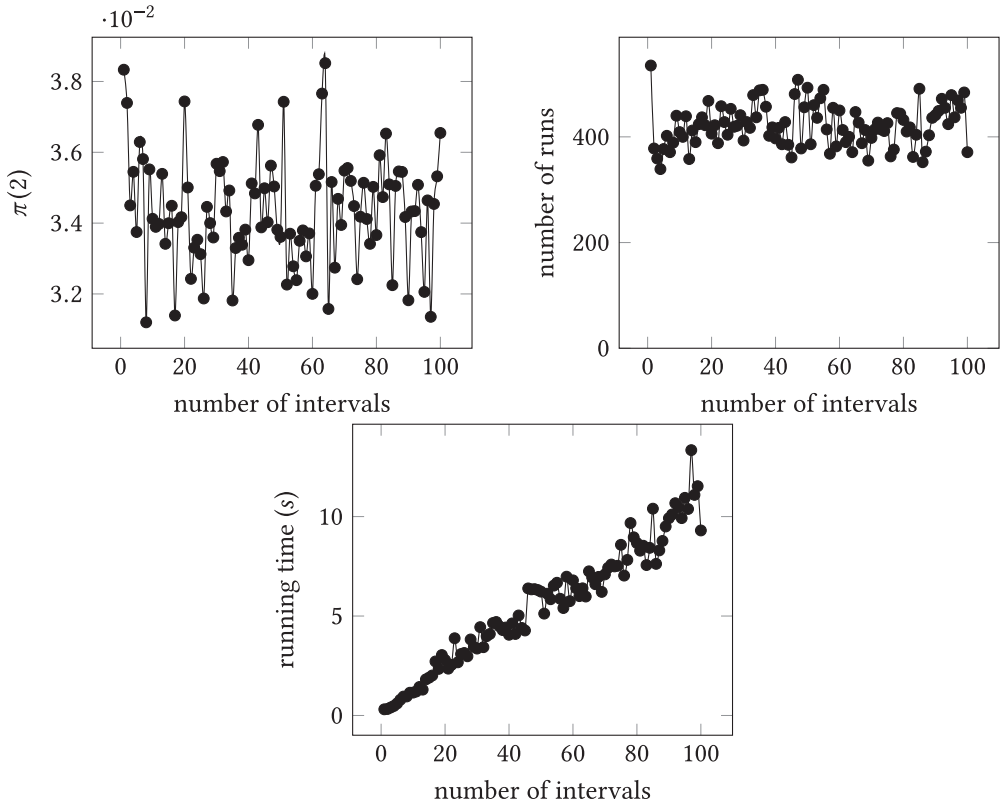

Fig. 6. These plots show the probability $\pi(2)$, the number of runs (with a maximum of 100,000) needed to get the desired accuracy, and the running time, respectively. The parameters that were used were $\mu = \binom{-1}{-0.5}$ and $\Sigma = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$. The lower and upper bound for the intervals were -5 and 2, respectively. The results show that the number of intervals does not seem to have any impact on the number of runs needed. The running time, however, seems to increase linearly as a function of the number of intervals.

give the results of two experiments; the difference is the level u that has to be reached (5 and 2, respectively). The results indicate that a negative covariance slows down both methods but also show that partitioned importance sampling is faster. When the correlation becomes positive, the opposite seems to hold.

6.4 Variable Number of Intervals

In this part, we restrict ourselves to partitioned importance sampling only. The goal is to find how the number of intervals affects the performance. From Figure 6, we can conclude that the number of intervals does not seem to have an effect on the number of trials needed. The total running time, however, does seem to suffer from a high number of intervals, though only linearly.

7 PROOF OF THEOREM 3.1

This section presents the proof of Theorem 3.1. The right-hand side of Equation (4) will be proved first as a lower bound and then as an upper bound for the left-hand side of Equation (4).

Lower Bound. First observe that $\pi(u) \geq \mathbb{P}(A_{su} > u, B_{tu} > u)$ for all s, t (where we allow ourselves, here and elsewhere, the imprecise notation su and tu when we mean their respective

rounded-off values). It thus follows, for all s and t, that

$$\liminf_{u\to\infty} \frac{1}{u} \ln \pi(u) \geq \liminf_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\mathscr{A}_{su}, \mathscr{B}_{tu}\right).$$

Since this holds for all s and t, we take the supremum over all s and t:

$$\liminf_{u\to\infty} \frac{1}{u} \ln \pi(u) \geq \sup_{s,t} \liminf_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\mathscr{A}_{su}, \mathscr{B}_{tu}\right).$$

Suppose, without loss of generality, that $s < t$ and let $\Delta > 0$. Then for all y, using the independence,

$$\begin{aligned}
\mathbb{P}\left(\mathscr{A}_{su}, \mathscr{B}_{tu}\right) &\geq \mathbb{P}\left(\frac{A_{su}}{su} > \frac{1}{s}, \frac{B_{su}}{su} \in [y, y+\Delta], \frac{B_{tu}}{tu} > \frac{1}{t}\right) \\
&\geq \mathbb{P}\left(\frac{A_{su}}{su} > \frac{1}{s}, \frac{B_{su}}{su} \in [y, y+\Delta], B_{tu} - B_{su} > u - suy\right) \\
&= \mathbb{P}\left(\frac{A_{su}}{su} > \frac{1}{s}, \frac{B_{su}}{su} \in [y, y+\Delta]\right) \cdot \mathbb{P}\left(B_{tu} - B_{su} > u - suy\right) \\
&= \mathbb{P}\left(\frac{A_{su}}{su} > \frac{1}{s}, \frac{B_{su}}{su} \in [y, y+\Delta]\right) \cdot \mathbb{P}\left(\frac{B_{(t-s)u}}{(t-s)u} > \frac{1-sy}{t-s}\right).
\end{aligned}$$

By Cramér's theorem, we have that the decay rate of this expression equals

$$-s \inf_{p>\frac{1}{s}, q\in[y,y+\Delta]} I(p,q) - (t-s) \inf_{q>\frac{1-sy}{t-s}} I_2(q).$$

As this relation holds for any Δ, using the continuity of I, we thus find that

$$\liminf_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\mathscr{A}_{su}, \mathscr{B}_{tu}\right) \geq -s \inf_{p>\frac{1}{s}} I(p,y) - (t-s) \inf_{q>\frac{1-sy}{t-s}} I_2(q).$$

This relation holds for any y, so in particular for all $y < 1/s$; as $I_2(q)$ increases in q for q positive, we obtain

$$\inf_{q>\frac{1-sy}{t-s}} I_2(q) = I_2\left(\frac{1-sy}{t-s}\right),$$

and, as a consequence,

$$\sup_{0<s<t} \liminf_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\mathscr{A}_{su}, \mathscr{B}_{tu}\right) \geq - \inf_{0<s<t, y<1/s} \left(s \inf_{p>\frac{1}{s}} I(p,y) + (t-s) I_2\left(\frac{1-sy}{t-s}\right)\right). \tag{10}$$

Now we put $v := t - s$ to obtain that the right-hand side of the previous display equals

$$\begin{aligned}
&- \inf_{s>0, y<1/s} s \left(\inf_{p>1/s} I(p,y) + (1-sy) \inf_{v>0} \frac{v}{1-sy} I_2\left(\frac{1-sy}{v}\right)\right) \\
&\quad = - \inf_{s>0, y<1/s} \left(s \inf_{p>\frac{1}{s}} I(p,y) + (1-sy)\alpha_2 \right) \\
&\quad = - \inf_{x>0, y<x} \left(\inf_{p>x} \frac{I(p,y)}{x} + \left(1 - \frac{y}{x}\right)\alpha_2 \right). \\
&\quad \geq - \inf_{x>0, y<x} \left(\frac{I(x,y)}{x} + \left(1 - \frac{y}{x}\right)\alpha_2 \right).
\end{aligned}$$

Upper Bound. Now we turn to the upper bound. We split the event in multiple sub-events. Fix some $s^\star$ and $t^\star$. Then the union bound implies that, for any $a > 0$,

$$\begin{aligned}\pi(u) &\leq \mathbb{P}\left(\exists s < (1+a)s^\star u : \mathscr{A}_s, \exists t < (1+a)t^\star u : \mathscr{B}_t\right)\\ &\quad + \mathbb{P}\left(\exists s \geq (1+a)s^\star u : \mathscr{A}_s, \exists t \in \mathbb{N} : \mathscr{B}_t\right)\\ &\quad + \mathbb{P}\left(\exists s \in \mathbb{N} : \mathscr{A}_s, \exists t \geq (1+a)t^\star u : \mathscr{B}_t\right)\\ &\leq \mathbb{P}\left(\exists s < (1+a)s^\star u : \mathscr{A}_s, \exists t < (1+a)t^\star u : \mathscr{B}_t\right)\\ &\quad + \mathbb{P}\left(\exists s \geq (1+a)s^\star u : \mathscr{A}_s\right) + \mathbb{P}\left(\exists t \geq (1+a)t^\star u : \mathscr{B}_t\right).\end{aligned}$$

We now show that the latter two terms have a higher decay rate (i.e., decay *faster*) than the first term. First, we use the union bound to get

$$\mathbb{P}\left(\exists s \geq (1+a)s^\star u : \mathscr{A}_s\right) \leq \sum_{s=(1+a)s^\star u}^{\infty} \mathbb{P}\left(A_s > u\right).$$

We first focus on the individual terms of the right-hand side. Using Markov's inequality, we get for all s:

$$\mathbb{P}\left(A_s > u\right) \leq \mathbb{E}(e^{\theta A_s})e^{-\theta u}.$$

First note that, since the process is a random walk,

$$\mathbb{E}(e^{\theta A_s}) = (\mathbb{E}(e^{\theta X}))^s = \Lambda_1(\theta)^s = e^{s\ln\Lambda_1(\theta)}.$$

Since this holds for all θ, we take the infimum:

$$\mathbb{P}\left(A_s > u\right) \leq \inf_\theta e^{s\ln\Lambda_1(\theta)}e^{-\theta u} = e^{-\sup_\theta(\theta u - s\ln\Lambda_1(\theta))} = e^{-sI_1(u/s)} \leq e^{-sI_1(0)};$$

in the second inequality we have used that $\mathbb{E}\left(X_1\right) < 0$. Now we return to the sum again. Using the display above, we get

$$\mathbb{P}\left(\exists s \geq (1+a)s^\star u : \mathscr{A}_s\right) \leq \sum_{s=(1+a)s^\star u}^{\infty} e^{-sI_1(0)} = \frac{e^{-I_1(0)\cdot(1+a)s^\star u}}{1-e^{-I_1(0)}},$$

so that for the decay rate of the probability above we get

$$\limsup_{u\to\infty}\frac{1}{u}\ln\mathbb{P}\left(\exists s \geq (1+a)s^\star u : \mathscr{A}_s\right) \leq -(1+a)s^\star I_1(0).$$

We conclude that the decay rate can be made arbitrarily large by letting $a \to \infty$. Obviously, the same procedure can be followed for $\mathbb{P}\left(\exists t \geq (1+a)t^\star u : \mathscr{B}_t\right)$. It thus follows that the first term has the lowest decay rate, and therefore the Principle of the Largest Term [8, Lemma 1.2.15] gives

$$\limsup_{u\to\infty}\frac{1}{u}\ln\pi(u) \leq \lim_{u\to\infty}\frac{1}{u}\ln\mathbb{P}\left(\exists s < (1+a)s^\star u : \mathscr{A}_s, \exists t < (1+a)t^\star u : \mathscr{B}_t\right). \tag{11}$$

Define $T := \max\{s^\star, t^\star\}$ and $\alpha := (1+a)T$. We introduce the scaled processes $\bar{A}_u(s) := \frac{1}{\alpha u}A_{\alpha u s}$ and $\bar{B}_u(t) := \frac{1}{\alpha u}B_{\alpha u t}$. The probability on the right-hand side above is then smaller than or equal to

$$\mathbb{P}\left(\sup_{s\leq 1}\bar{A}_u(s) \geq 1/\alpha, \sup_{t\leq 1}\bar{B}_u(t) \geq 1/\alpha\right).$$

Let f be the sample path of a two-dimensional function. Also, let $\phi_1(f) := \sup_{s\le 1} f_1(s)$, and likewise let $\phi_2(f) := \sup_{t\le 1} f_2(t)$. We now invoke Mogulskii's theorem (see Reference [8, Theorem 5.1.2]). This gives us

$$-\inf_{\{f:\phi_1(f)\ge 1/\alpha,\phi_2(f)\ge 1/\alpha\}^{\circ}} J(f) \le \liminf_{u\to\infty} \frac{1}{\alpha u} \ln \mathbb{P}\left(\sup_{s\le 1} \bar{A}_u(s) \ge 1/\alpha, \sup_{t\le 1} \bar{B}_u(t) \ge 1/\alpha\right)$$
$$\le \limsup_{u\to\infty} \frac{1}{\alpha u} \ln \mathbb{P}\left(\sup_{s\le 1} \bar{A}_u(s) \ge 1/\alpha, \sup_{t\le 1} \bar{B}_u(t) \ge 1/\alpha\right) \le -\inf_{\{f:\phi_1(f)\ge 1/\alpha,\phi_2(f)\ge 1/\alpha\}^{c}} J(f),$$

with

$$J(f) = \int_0^1 I(f'(t))\,\mathrm{d}t;$$

note that the conditions imposed allow that Mogulskii's theorem can be applied. Note that the set over which the infimum is taken is closed, hence we will drop the closure operator. So an upper bound for the right-hand side of Equation (11) is

$$-\alpha \inf_{\{f:\phi_1(f)\ge 1/\alpha,\phi_2(f)\ge 1/\alpha\}} J(f). \tag{12}$$

In the calculations below, we will drop the factor α in front of the infimum; we will later see that it cancels.

Assume that $\bar{A}_u(\cdot)$ hits u for the first time at time $\bar{s}$ and $\bar{B}_u(\cdot)$ hits u for the first time at time $\bar{t}$, i.e., $\bar{s} \equiv \bar{s}(f) = \inf_{s\in[0,1]}\{s : f_1(s) \ge 1/\alpha\}$ and likewise for $\bar{t}$. We can then rewrite the upper bound in two cases:

$$\inf_{\{f:\phi_1(f)\ge 1/\alpha,\phi_2(f)\ge 1/\alpha\}} J(f) = \inf_{\{f:\bar{s}(f)\le 1,\bar{t}(f)\le 1\}} J(f)$$
$$= \min\left\{\inf_{\{f:\bar{s}(f)\le 1,\bar{t}(f)\le 1,\bar{s}\le\bar{t}\}} J(f), \inf_{\{f:\bar{s}(f)\le 1,\bar{t}(f)\le 1,\bar{s}>\bar{t}\}} J(f)\right\}.$$

We will now focus on the first entry of the minimum above; the second entry can be treated analogously. It can be rewritten as

$$\inf_{v\le 1/\alpha} \inf_{\{f:\bar{s}(f)\le 1,\bar{t}(f)\le 1,\bar{s}\le\bar{t},f_2(\bar{s})=v\}} J(f).$$

Pick a fixed but arbitrary f which is confined to the restrictions in the infima above. We wil now rewrite $J(f)$ as a sum of three integrals:

$$J(f) = \int_0^1 I(f'(t))(f'(t))\,\mathrm{d}t = \int_0^{\bar{s}} I(f'(t))\,\mathrm{d}t + \int_{\bar{s}}^{\bar{t}} I(f'(t))\,\mathrm{d}t + \int_{\bar{t}}^1 I(f'(t))(f'(t))\,\mathrm{d}t.$$

In the spirit of Reference [11], we will construct a straightened path $\tilde{f}$ and then show that the upper bound in the LDP is the same as $-J(\tilde{f})$. Let

$$\left(\tilde{f}_1'(\tau), \tilde{f}_2'(\tau)\right) = \begin{cases} \left(\frac{1}{\alpha\bar{s}}, \frac{v}{\bar{s}}\right) & \text{if } 0 \le \tau \le \bar{s}; \\ \left(\frac{c^\star}{\bar{t}-\bar{s}}, \frac{1/\alpha - v}{\bar{t}-\bar{s}}\right) & \text{if } \bar{s} < \tau \le \bar{t}; \\ (\mu, \nu) & \text{if } \bar{t} < \tau \le 1, \end{cases}$$

where $c^\star := \arg\min_c I\left(\frac{c}{\bar{t}-\bar{s}}, \frac{1/\alpha - v}{\bar{t}-\bar{s}}\right)$, $\mu := \mathbb{E}(X_1)$ and $\nu := \mathbb{E}(Y_1)$. Now note that

(1) using Jensen's inequality,

$$\int_0^{\bar{s}} I(\tilde{f}'(t))\,\mathrm{d}t = \bar{s} I\left(\frac{1}{\alpha\bar{s}}, \frac{v}{\bar{s}}\right) = \bar{s} I\left(\frac{1}{\bar{s}}\int_0^{\bar{s}} f'(t)\,\mathrm{d}t\right) \le \int_0^{\bar{s}} I(f'(t))\,\mathrm{d}t;$$

(2) using the minimisation gives us

$$\int_{\bar{s}}^{\bar{t}} I(\tilde{f}'(t))\,\mathrm{d}t = (\bar{t}-\bar{s})I\left(\frac{c^\star}{\bar{t}-\bar{s}}, \frac{1/\alpha - v}{\bar{t}-\bar{s}}\right) \leq (\bar{t}-\bar{s})I\left(\frac{1}{\bar{t}-\bar{s}}\int_{\bar{s}}^{\bar{t}} f'(t)\,\mathrm{d}t\right) \leq \int_{\bar{s}}^{\bar{t}} I(f'(t))\,\mathrm{d}t;$$

(3) using Reference [11, Lemma 2.6.iv] gives us

$$\int_{\bar{t}}^{1} I(\tilde{f}'(t))\,\mathrm{d}t = 0 \leq \int_{\bar{t}}^{1} I(f'(t))\,\mathrm{d}t,$$

and hence, $J(f) \geq J(\tilde{f})$. Furthermore,

$$\inf_{v\leq 1/\alpha}\inf_{\{f:\bar{s}(f)\leq 1,\bar{t}(f)\leq 1,\bar{s}\leq\bar{t},f_2(\bar{s})=v\}} J(f) \leq \inf_{v\leq 1/\alpha}\inf_{\{\tilde{f}:\bar{s}(\tilde{f})\leq 1,\bar{t}(\tilde{f})\leq 1,\bar{s}\leq\bar{t},\tilde{f}_2(\bar{s})=v\}} J(\tilde{f}),$$

where on the right-hand side we restrict the infimum to straightened paths as described above. Hence, this inequality is in fact an equality.

We will now focus on $J(\tilde{f})$. Note that it is equal to

$$\bar{s}I\left(\frac{1}{\alpha\bar{s}}, \frac{v}{\bar{s}}\right) + (\bar{t}-\bar{s})I\left(\frac{c^\star}{\bar{t}-\bar{s}}, \frac{1/\alpha - v}{\bar{t}-\bar{s}}\right).$$

When we now bring back the infimum, we get

$$\inf_{v\leq 1/\alpha, 0\leq\bar{s}\leq\bar{t}\leq 1} \bar{s}I\left(\frac{1}{\alpha\bar{s}}, \frac{v}{\bar{s}}\right) + (\bar{t}-\bar{s})I\left(\frac{c^\star}{\bar{t}-\bar{s}}, \frac{1/\alpha - v}{\bar{t}-\bar{s}}\right),$$

or, using the definition of $c^\star$,

$$\inf_{v\leq 1/\alpha, 0\leq\bar{s}\leq\bar{t}\leq 1} \bar{s}I\left(\frac{1}{\alpha\bar{s}}, \frac{v}{\bar{s}}\right) + (\bar{t}-\bar{s})\inf_c I\left(\frac{c}{\bar{t}-\bar{s}}, \frac{1/\alpha - v}{\bar{t}-\bar{s}}\right).$$

This is bigger than or equal to

$$\inf_{v\leq 1/\alpha, 0\leq\bar{s}\leq\bar{t}\leq 1} \bar{s}I\left(\frac{1}{\alpha\bar{s}}, \frac{v}{\bar{s}}\right) + (\bar{t}-\bar{s})\inf_{c\in\mathbb{R}, z>0} \frac{I\left(\frac{c}{\bar{t}-\bar{s}}, z\right)}{z\frac{\bar{t}-\bar{s}}{1/\alpha - v}},$$

and cancelling the factors gives us

$$\inf_{v\leq 1/\alpha, 0\leq\bar{s}\leq\bar{t}\leq 1} \bar{s}I\left(\frac{1}{\alpha\bar{s}}, \frac{v}{\bar{s}}\right) + \left(\frac{1}{\alpha} - v\right)\inf_{c\in\mathbb{R}, z>0} \frac{I\left(\frac{c}{\bar{t}-\bar{s}}, z\right)}{z}.$$

This is again bigger than or equal to

$$\inf_{v\leq 1/\alpha, 0\leq\bar{s}\leq\bar{t}\leq 1} \bar{s}I\left(\frac{1}{\alpha\bar{s}}, \frac{v}{\bar{s}}\right) + \left(\frac{1}{\alpha} - v\right)\inf_{z>0} \frac{I_2(z)}{z}.$$

Now define $x := 1/\alpha\bar{s}$ and $y := \alpha v x$. Then the expression above is equal to

$$\inf_{v\leq 1/\alpha, 0\leq\bar{s}\leq\bar{t}\leq 1, x=1/\alpha\bar{s}, y=\alpha v x} \frac{1}{\alpha x} I(x,y) + \left(\frac{1}{\alpha} - \frac{y}{\alpha x}\right)\inf_{z>0} \frac{I_2(z)}{z},$$

which equals

$$\frac{1}{\alpha}\left(\inf_{x\geq 1/\alpha, y\leq x} \frac{1}{x} I(x,y) + \left(1 - \frac{y}{x}\right)\inf_{z>0} \frac{I_2(z)}{z}\right),$$

which majorises

$$\frac{1}{\alpha}\left(\inf_{x>0,\,y\le x}\frac{1}{x}I(x,y)+\left(1-\frac{y}{x}\right)\inf_{z>0}\frac{I_2(z)}{z}\right).$$

Recall that the factor $\frac{1}{\alpha}$ cancels against the factor α of Equation (12), and, hence, we have proven the upper bound.

8 CONCLUDING REMARKS

In this article, we studied both logarithmic asymptotics and several numerical methods to study large delay probabilities of two correlated queues. In the first part of the article, the first main result, Theorem 3.1, was given, which provided an expression for the decay rate of the probability of both components ever reaching some high level. The second part consisted of analysing two numerical procedures, namely a naive importance sampling procedure and partitioned importance sampling. It was shown that the former method is not necessarily asymptotically optimal. This is caused by the undershoot of the slowest component. This problem is overcome by the second method, partitioned importance sampling. It was indeed shown that this procedure is asymptotically optimal. Subsequently, numerical results of simulation experiments were shown, confirming the theory.

REFERENCES

[1] E. S. Badila. 2015. *Queues and Risk Models*. Ph.D. thesis, Technische Universiteit Eindhoven.
[2] Jose Blanchet. 2013. Optimal sampling of overflow paths in Jackson networks. *Math. Operat. Res.* 38,4 (2013), 698–719.
[3] Jose Blanchet and Henry Lam. 2012. State-dependent importance sampling for rare-event simulation: An overview and recent advances. *Surv. Operat. Res. Manage. Sci.* 17,1 (2012), 38–59.
[4] José Blanchet and Michel Mandjes. 2009. Rare event simulation for queues. In *Rare Event Simulation Using Monte Carlo Methods*, Gerardo Rubino and Bruno Tuffin (eds.), chapter 5. John Wiley & Sons.
[5] Ewan Jacov Cahen, Michel Mandjes, and Bert Zwart. 2017. Rare event analysis and efficient simulation for a multi-dimensional ruin problem. *Probability in the Engineering and Informational Sciences*, 1–19.
[6] Pieter-Tjerk de Boer. 2006. Analysis of state-independent importance-sampling measures for the two-node tandem queue. *ACM Trans. Model. Comput. Simul.* 16, 3 (2006), 225–250.
[7] K. Dębicki, A. B. Dieker, and T. Rolski. 2007. Quasi-product forms for Lévy-driven fluid networks. *Math. Operat. Res.* 32, 3 (2007), 629–647.
[8] A. Dembo and O. Zeitouni. 1998. *Large Deviations Techniques and Applications* (2nd ed.). Springer-Verlag, 1998.
[9] Paul Dupuis and Hui Wang. 2009. Importance sampling for Jackson networks. *Que. Syst.* 62, 1–2 (2009):, 113–157.
[10] Paul Dupuis, Ali Devin Sezer, and Hui Wang. 2007. Dynamic importance sampling for queueing networks. *Ann. Appl. Probab.* 17, 4 (2007), 1306–1346.
[11] A. Ganesh, N. O'Connell, and D. Wischik. 2004. *Big Queues*. Springer.
[12] Philip Heidelberger. 1995. Fast simulation of rare events in queueing and reliability models. *ACM Trans. Model. Comput. Simul.* 5, 1 (1995), 43–85.
[13] Sandeep Juneja and Perwez Shahabuddin. 2006. Chapter 11 rare-event simulation techniques: an introduction and recent advances. *Handbooks in Operations Research and Management Science*, Shane G. Henderson and Barry L. Nelson (Eds.). Vol. 13. Elsevier, 291–350.
[14] O. Kella. 1993. Parallel and tandem fluid networks with dependent Lévy inputs. *Ann. Appl. Probab.* 3, 3 (1993), 682–695.
[15] Offer Kella and Ward Whitt. 1992. A tandem fluid network with Lévy input. In *Queueing and related models*, U. N. Bhat and I. V. Basawa (eds.), chapter 7. Oxford University Press, 112–128.
[16] Pierre L'Ecuyer, Jose H. Blanchet, Bruno Tuffin, and Peter W. Glynn. 2010. Asymptotic robustness of estimators in rare-event simulation. *ACM Trans. Model. Comput. Simul.* 20, 1 (2010).
[17] Pascal Lieshout and Michel Mandjes. 2007. Tandem brownian queues. *Math. Methods Operat. Res.* 66, 2 (2007), 275–298.
[18] Michel Mandjes. 2004. Packet models revisited: Tandem and priority systems. *Que. Syst.* 47, 4 (2004), 363–377.
[19] John S. Sadowsky. 1991. Large deviations theory and efficient simulation of excessive backlogs in a $GI/GI/m$ queue. *IEEE Trans. Autom. Control* 36, 12 (1991), 1383–1394.
[20] David Siegmund. 1976. Importance sampling in the Monte Carlo study of sequential tests. *Ann. Stat.* 4, 4 (1976), 673–684.

Received December 2016; revised November 2017; accepted November 2017